5-12-2020

Free Speech in the Digital Age: Deepfakes and the Marketplace of Ideas

Suyoung Baek

Free Speech in the Digital Age: Deepfakes and the Marketplace of Ideas

Abstract

The threat of deepfakes is well-documented in the existing literature. Deepfake technology has emerged as a powerful tool with which vulnerable individuals could easily become targets of novel forms of exploitation and sabotage. Additionally, deepfakes' unique capacity to distort people's sense of reality exacerbates truth decay. The growing influence of social media and our deep-rooted cognitive biases further escalate the harms of deepfakes. Despite these apparent concerns, scholars have noted that the regulation of deepfakes confronts a constitutional challenge in the American context, stemming from Supreme Court cases such as New York Times v. Sullivan and U.S. v. Alvarez. In both cases, the Court emphasized the importance of protecting false speech on the grounds that it constitutes an integral part of the "marketplace of ideas." This paper aims to show how the broad range of harms posed by deepfakes in the digital age calls for a departure from employing the Times and Alvarez approaches to assessing the constitutionality of deepfakes.

Keywords

philosophy, deepfakes, marketplace of ideas, free speech

Disciplines

First Amendment | Philosophy

FREE SPEECH IN THE DIGITAL AGE:

DEEPFAKES AND THE MARKETPLACE OF IDEAS

by

Suyoung Baek

Submitted to the

Department of Philosophy, Politics, and Economics

University of Pennsylvania

Thesis Advisor: Dr. Samuel Freeman

May 12, 2020

ABSTRACT

The threat of deepfakes is well-documented in the existing literature. Deepfake technology has emerged as a powerful tool with which vulnerable individuals could easily become targets of novel forms of exploitation and sabotage. Additionally, deepfakes' unique capacity to distort people's sense of reality exacerbates truth decay. The growing influence of social media and our deep-rooted cognitive biases further escalate the harms of deepfakes. Despite these apparent concerns, scholars have noted that the regulation of deepfakes confronts a constitutional challenge in the American context, stemming from Supreme Court cases such as *New York Times v. Sullivan* and *U.S. v. Alvarez.* In both cases, the Court emphasized the importance of protecting false speech on the grounds that it constitutes an integral part of the "marketplace of ideas." This paper aims to show how the broad range of harms posed by deepfakes in the digital age calls for a departure from employing the *Times* and *Alvarez* approaches to assessing the constitutionality of deepfakes.

Introduction

The word “deepfake” is a portmanteau of “deep learning” and “fake” (Rouse, 2020). It refers to a type of artificial intelligence (AI) technology that incorporates a machine learning technique called generative adversarial networks (GANs) (Rouse, 2020). GANs was first introduced in 2014 by Ian Goodfellow and other researchers at the University of Montreal (Goodfellow et al, 2014). The idea is to use a pair of neural networks – one of which is called the “generator,” and the other, the “discriminator” – to synthesize artificial media or multimedia content that is indistinguishable from its authentic counterpart (Brownlee, 2019). One of the most striking features of this algorithmic architecture is its ability to use as little as one image of a person to create a video clip of that person saying or doing things they never said or did in real life (Libby, 2019).

In recent years, deepfake technology has earned its reputation as a threat to our already vulnerable information ecosystem (Schwartz, 2018). Until late 2017, the use of this machine learning technique was mostly confined to the area of AI research (Schwartz, 2018). It was only when a Reddit user who, under the moniker “Deepfakes,” began posting digitally altered pornographic videos in which celebrities’ faces were superimposed onto the bodies of women in pornographic movies, that this technology became widely known in the public domain (Schwartz, 2018). By the time Reddit later banned the posting and dissemination of deepfakes from its platform, the creator of the videos had released “FakeApp,” an easy-to-use platform for making forged media (Schwartz, 2018). With the help of FakeApp, deepfake technology became widely known

and available to the public, resulting in a dramatic increase in the number of individuals who utilized this technology to generate and disseminate deepfakes online, mainly through social media platforms (Schwartz, 2018). In September 2019, the AI firm Deeptrace found approximately 15,000 deepfake videos online, 96% of which were pornographic (Sample, 2020).

The goal of this paper is to provide an in-depth assessment of this disruptive technology in order to construct a more robust framework for evaluating whether deepfakes should be protected under the category of "false speech." The paper proceeds as follows: Part I discusses the ways in which deepfakes present an unprecedented threat to individuals and to society. Part II explains how false speech has come to be viewed as a form of speech that warrants a degree of protection and why deepfakes, due to the novel threats they pose, should not fall under this particular category of speech. Part III discusses the importance of distinguishing malicious deepfakes from satire and parody in that the latter two are legally permissible and socially valuable types of speech, while malicious deepfakes assume deliberate deception. Finally, Part IV summarizes some of the legal and constitutional challenges that need to be overcome in order to bring about constructive and lasting changes with regard to the looming threat of deepfakes.

PART I: THE PROBLEM OF DEEPFAKES

Why are deepfakes dangerous?

In response to the extensive use of deepfake technology in the realm of pornography and increasingly elsewhere, many scholars, especially in the areas of law and policy, have voiced their concerns about the potential for deepfakes to become powerful mechanisms to exploit and sabotage others, as well as to harm society by disrupting democratic discourse on important policy questions (Libby, 2019). Moreover, recent events, such as the Russian intervention in the 2016 presidential election, combined with the growing risk of cyberwar escalation, have placed the issue of evaluating the danger of deepfakes at the top of many organizations' agendas, including those of some of the largest social media networks, such as Twitter and Facebook (Ghaffary, 2020; Romm, Harvwell, Stanley-Becker, 2020).

One of the most seminal works in this area of concern is a paper titled "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security," written by Danielle Citron, a professor of law at Boston University, and her colleague Bobby Chesney. In the paper, Citron and Chesney (2018) explore a number of ways in which deepfakes could cause harm not only to individuals but also to society at large. With an emphasis on the findings of Citron and Chesney (2018), the rest of Part I aims to (1) summarize the general view among scholars on the danger of deepfake technology, (2) discuss the role of social media and confirmation bias in enhancing the danger of misinformation, especially during a time of crisis, and (3) introduce how the issue of free

speech presents constitutional challenges against efforts to regulate the circulation of deepfakes.

Citron and Chesney (2018) begin their discussion on the many harms of deepfake technology by focusing on deepfakes' capacity to serve as powerful mechanisms to exploit and sabotage others. In particular, they warn of the possibility of deepfakes escalating the severity of rape threats by giving off the impression that victims can be sexually abused at whim (Citron and Chesney, 2018). Even more problematic, the targets of various forms of nonconsensual pornography made possible by deepfakes tend to be disproportionately female as well as disproportionately queer, rendering such types of digital manipulation especially dangerous to some of the most vulnerable groups within society (Franks and Waldman, 2019). The authors further argue that even if a deepfake video involves no sexual violence, any kind of abuse illustrated can be exploited to threaten, intimidate, and inflict psychological harm on the individual depicted or those who care for that individual's safety (Citron and Chesney, 2018). Citron and Chesney (2018) also examine how deepfake technology can be used to tarnish a person's reputation by portraying the person in a negative light in the presence of his or her rivals. For instance, a malicious agent could falsely implicate a person by creating fake evidence using deepfake technology. The agent could then make the depicted person sustain serious damage, whether it be reputational, financial, or psychological, by sharing the deepfake with the person's rivals (Citron and Chesney, 2018).

In addition to describing the extent to which deepfake technology can harm individuals through exploitation and sabotage, Citron and Chesney (2018) discuss the

distinct possibility of deepfakes causing considerable harm to society. To begin with, they argue that the threat posed by deepfakes at a societal level involves systemic dimensions, for the damage caused by deepfakes may undermine efforts to maintain the wellbeing of democratic institutions. They point to the intervention of the Russian government in the 2016 U.S. presidential election as compelling evidence of how foreign actors can seriously destabilize democratic discourse by exploiting the capacity of deepfakes to distort factual information and thus manipulate beliefs – on the grounds that one of the key ingredients for sustainable democratic discourse is "a shared universe of facts and truths" backed by empirical evidence (Citron and Chesney, 2018).

Yet, the level of harm deepfakes pose to society extends beyond merely disrupting democratic discourse. When debates on questions of policy become infused with deliberate falsehoods transmitted by malicious agents, voters' ability to make informed, rational decisions about candidates running for an office of any sort is unequivocally damaged. On this point, Manzi (2019) suggests that voters' susceptibility to falsities disrupts their ability to choose candidates who represent their interests. In the current legal atmosphere, whether such disruptions amount to a legitimate case of electoral fraud or vote rigging would be difficult to demonstrate, and it would be even more difficult to bring the actors involved to justice – as we saw in the case of the Russian intervention. Another reason assigning commensurate responsibility in this regard presents challenges is that the constitutionality of false speech in the context of the First Amendment has long been regarded as a highly contextual matter by the courts – a consideration that will be elaborated upon in Part II. Nonetheless, the notion that

deepfakes, by disrupting people's sense of reality, have the capacity to influence public opinion and thus destabilize our democratic institutions seems highly plausible and therefore warrants further scrutiny.

The role of social media and confirmation bias

When discussing the dangers associated with deepfakes, it is also important to understand how the nature of today's communication environment, combined with the effects of cognitive bias on how people perceive information of unknown accuracy, enhances the capacity of deep fakes to cause serious harm. With respect to today's communication environment, for instance, we are seeing a sharp decline in traditional media such as newspapers, television, and radio in favor of social media as the main source of news (Manzi, 2019). According to a survey conducted by the Pew Research Center in 2019, 55% of U.S. adults get their news from social media either "often" or "sometimes," and 88% of Americans recognize that social media companies have some control over the mix of news that people consume each day (Suciu, 2019). The reason such a shift towards social media presents challenges with respect to the circulation of deepfakes is that social media networks, unlike traditional news organizations, do not operate under the principle of safeguarding democracy or enabling truth-seeking in the ideas of the marketplace (Manzi, 2018). It has been argued that the attention-grabbing algorithms underlying social media play a significant role in propelling malicious actors to inject social media networks with misinformation to sow confusion, political discord, prejudice, and chaos (Deibert, 2019). Fournier (2020), for example, writes that social media has enabled foreign intelligence agencies and entities to "covertly inject

disinformation through the use of highly searched hashtags, keywords, and precise timing."

The power of social media in facilitating the spread of misinformation was recently underscored by the novel COVID-19 outbreak. Alarmed by the amount of harmful memes and misinformation related to the outbreak on social media platforms like Twitter and Facebook, Michigan State University's Communication Arts and Sciences professors conducted a close analysis of the situation. They concluded that the outbreak has caused an "Infodemic" and that social media in particular has created conditions ripe for misinformation to prevail (Priebe, 2020). Professor Dustin Carnahan attributes this phenomenon to some of social media's most prominent features. He suggests that because interactions on social media usually occur between people who have personal connections or among those who admire or respect one another, people are less likely to question the believability of what they see (Priebe, 2020). He further claims that the viral and unfiltered nature of social media enables information to spread without extensive scrutiny by credible sources.

One notable example of misinformation related to COVID-19 involves President Trump's passing comment during a press briefing about using ultraviolet light inside the human body or a disinfectant by "injection" as a treatment for COVID-19 (Panetta, 2020). Because social media is often used to quickly pass along statements made by public officials or well-known celebrities, even if doing so risks the dissemination of misinformation to a sizable audience, President Trump's rather "spontaneous" suggestion was instantly shared through various social media platforms. Although President Trump

phrased his suggestion as "it'd be interesting to check"[1] rather than a clear recommendation based on science and facts, the extraordinary rate at which social media users were sharing this comment caused a large number of people to falsely believe that the President was telling people to inject themselves with bleach or isopropyl alcohol (Mahadevan, 2020). Although scientific experts and medical professionals were quick to repudiate the efficacy of such courses of action, within hours of the President's suggestion, there were already reports indicating a spike in New Yorkers ingesting household cleaners, demonstrating the power of information, let alone misinformation, coming from public figures with considerable influence (Sanders & Sommerfeldt, 2020).

During emergency situations, people have a natural tendency to rely on authoritative knowledge when forming judgments on how to protect themselves from potential dangers (Petryna, 2002). When people's susceptibility to become less prudent in their judgment in the face of uncertainty is enmeshed with a disruptive technology that enables individuals to become subjugated to a false sense of reality, the extent of the harm deepfakes can pose grows exponentially. As an illustration, consider the hypothetical scenario in which a malicious agent decides to employ deepfake technology to create and disseminate fake videos of political leaders or well-known medical professionals making dangerous suggestions with respect to COVID-19 – perhaps to sabotage the reputations of those portrayed in the deepfakes or to sow distrust, confusion, and panic among the public. Under this scenario, the unprecedented pace at which

[1] Trump's exact words during the April 23, 2020 coronavirus briefing: "I see the disinfectant, where it knocks it out in one minute…And is there a way we can do something like that by injection inside or almost a cleaning because you see it gets in the lungs and it does a tremendous number on the lungs, so it'd be interesting to check that so that you're going to have to use medical doctors with, but it sounds interesting to me" (Mahadevan, 2020).

misinformation disseminates, coupled with the possibility of such misinformation resulting in life-threatening decisions, could introduce a magnitude of harm that would be neither quantifiable nor controllable.

Given that, as of December 2019, there are approximately 3 billion active social media users, which amounts to almost half of the entire world population, it is reasonable to speculate that deepfakes' unique capacity to blur the line between fact and fiction can cause significant harm. This is particularly true in the domains of health and science where conflicting claims and mixed messaging can expose people to a wide range of risks, as evidenced by President Trump's *suggestion* about injecting disinfectants as treatment for COVID-19 rapidly turning into a *recommendation*, unbeknown to the speaker himself, after a series of reposts and spins made possible by the viral and unfiltered nature of social media. (Clement, 2020). Another factor that makes the unchecked proliferation of deepfakes especially dangerous is people's susceptibility to information that align with their preexisting beliefs, a tendency known as "confirmation bias" (Casad, 2019). The idea is that people have difficulty processing information in a rational, unbiased manner once they have formed an opinion about an issue (Casad, 2019). In other words, when presented with information of unknown accuracy, an individual will generally respond based on his or her preconceived notions rather than on empirical evidence (Sunstein, 2019). Cass Sunstein, a professor at Harvard Law School, argues that the effects of confirmation bias are especially dangerous when the information presented is false in that "once a false rumor has thus been accepted, a correction then becomes difficult to accept" (Sunstein, 2019). In the same vein, Erza

Waldman (2018), a professor at New York Law School, contends that false information can harden biases, which increases polarization. This has the effect of eroding trust in traditional reporting and encouraging the selection of information that confirms one's biases. Considering how well deepfakes can distort reality by generating content with real people saying and doing things they never said or did, the pervasive effects of confirmation bias paint a grim picture of the scope of harm deepfakes could cause.

The issue of free speech

At first glance, the costs of allowing deepfakes to freely roam around social media platforms may seem to justify some degree of regulation to prevent further harm. Upon further examination, however, one becomes cognizant of the broader significance of deepfakes as a form of "false speech" whose constitutionality is still being debated among legal scholars. Part II seeks to explain how false speech has come to be recognized as a form of speech that warrants a degree of protection under the First Amendment. Relevant Supreme Court cases will be reviewed, followed by a discussion on how deepfakes, due to their unprecedented capacity to undermine free debates and manipulate views that are critical to maintaining the democratic value of elections, ultimately call for a different, less conventional approach to understanding the constitutional value of false speech, especially *harmful* false speech, in the digital age.

PART II: DEEPFAKES AND THE FIRST AMENDMENT

The free speech argument *against* the regulation of deepfakes

The most frequently used reasoning against the regulation of deepfakes is that they, despite their inherent falsity, implicate freedom of expression (Hall, 2018). In fact, despite broad consensus on the need to address deepfakes and their growing influence, many scholars have been keen to point out the possibility of such efforts going too far. For example, Sharon Bradford Franklin, Policy Director for New America's Open Technology Institute warned that we must "avoid establishing legal rules that will push too far in the opposite direction, and engage in censorship of free expression online" (Franklin, 2019). Similarly, David Greene, the Electronic Frontier Foundation's civil liberties director, said that California's new political deepfake law, which makes it illegal for individuals to post manipulated content that give a "false impression of a political candidate's actions or words" in the 60 days before an election, is "overbroad, vague, and subjective" and fails to strike an appropriate balance between preventing harm and protecting the value of free speech[2] (Fischer, 2019). Tsukayama, McKinney, and Williams (2019) also advised against rushing to regulate deepfakes, claiming that while it is important for society to acknowledge the harmful uses of deepfakes and hold the people who cause them liable for their behaviors, it needs to do so in a way that does not censor lawful and socially valuable speech – a point of contention that will be further examined in Part III.

---

[2] In October 2019, California Governor Gavin Newsom signed two deepfake bills into state law. The first legislation makes it illegal to post manipulated videos and pictures that give a "false impression of a political candidate's actions or words" in the 60 days before an election. The second legislation allows residents to sue anyone who uses deepfake technology to place them in pornographic material without their consent (Fischer, 2019).

The value of false speech under "the marketplace of ideas"

Even though the Supreme Court has yet to evaluate the constitutionality of regulating deepfakes specifically, the Court has made several decisions involving the issue of "false speech" – granted, with some inconsistencies (Chemerinsky, 2018). For example, *NYT v. Sullivan* (1964) established that false political speech enjoys constitutional protection insofar as its prohibition would chill truthful speech, rendering the regulation of false speech particularly challenging. Delivering for the majority opinion of the Court, Justice William Brennan wrote that the decision was predicated upon "the background of a profound national commitment to the principle that debate on public issues should be uninhibited, robust, and wide-open" (NYT v. Sullivan, 1964). The Court explained that false statements, in this regard, are inevitable and thus must be protected if freedoms of expression are to have the "breathing space" that they need to survive (NYT v. Sullivan, 1964). Moreover, Justice Brennan wrote that defendants can only be required to pay damages to a public official for libel if a plaintiff is able to show that there was "actual malice," – that is, knowledge that the claim was false or reckless disregard as to the falsity of the statement (NYT v. Sullivan, 1964). The court, by emphatically rejecting that falsity alone suffices as a basis to deny First Amendment protection, especially when the speech in question is a political one, established a powerful precedent with respect to how false political speech should be viewed in the context of First Amendment law.

The Court's reference to the significance of an "uninhibited, robust, and wide-open debate" embodies one of the most influential governing principles in First

Amendment law: the "marketplace of ideas." Conceived by John Stuart Mill and later incorporated by Justice Oliver Wendell Holmes in his dissenting opinion in *Abrams v. United States* (1919), the marketplace of ideas is the belief that "the test of the truth or acceptance of ideas depends on their competition with one another and not on the opinion of a censor, whether one provided by the government or by some other authority" (Hudson, 2017). In line with this reasoning, the court has come to view false speech as providing a necessary condition under which true opinions can ascertain "the truth" (Manzi, 2019). *U.S. v. Alvarez (2012)* is a more recent case in which the Court utilized the marketplace of ideas metaphor to place false opinions under the category of protected speech under the First Amendment. Justice Anthony Kennedy, writing for the plurality opinion, concluded that false statements can be regulated only to the extent that defendants intend to cause "legally cognizable harm" and that a direct causal link exists between the "restriction imposed and the injury to be prevented" (U.S. v. Alvarez, 2012). The value of false opinions was highlighted once again in Susan B. Anthony List v. Driehaus (2014), in which the Court recognized the harms of an Ohio law that criminalized making false statements about candidates during political campaigns.

Despite the Court's repeated emphasis on the value of false speech in the context of the marketplace of ideas, the Court has also argued that false statements "are not protected by the First Amendment in the same manner as truthful statements" (Brown v. Hartlage, 1982). In Hustler Magazine, Inc. v. Falwell (1988), the Court held that "false statements of fact are particularly valueless [because] they interfere with the truth-seeking function of the marketplace of ideas." A similar view was articulated in the

Court's earlier Gertz v. Robert Welch, Inc. (1974) decision, in which it asserted that no false statement has constitutional value in that "neither the intentional lie nor the careless error materially advances society's interest in 'uninhibited, robust, and wide-open' debate on public issues." Chemerinsky (2018) explains that the apparent inconsistency in how the court has dealt with cases related to false speech is inevitable, because any analysis regarding false speech "must be contextual" and must reflect the "balancing of competing interests." The Court's *actual malice* standard, which gives false political speech greater protection in defamation cases, exemplifies how the Court has come to engage in a contextual analysis of determining whether a speaker's statement should be protected, despite its falsity, under the First Amendment (New York Times v. Sullivan, 1964).

The limitations of conventional free speech discourse

When one examines the ways in which the Supreme Court has constructed and redefined the American tradition of free speech, especially regarding censorship, it is evident that deepfakes do not fall nicely into the categories of speech that the court has deliberated over throughout history (Kalven, 1988). Because deepfake technology is a relatively new invention of the 21st century and its true impact is yet to be revealed, it is important to understand how deepfakes differ from other types of protected and unprotected categories of speech. This evaluation will provide helpful reference points from which we can assess the ability of deepfakes to enhance or undermine the truth-seeking function of the marketplace of ideas.

Responding to the assertion that deepfakes fall under the category of "false speech," mainly in the context of *NYT v. Sullivan* and *U.S. v. Alvarez*, and thus warrant

constitutional protection as a valid form of false speech, the first two sections of Part II have illustrated how false speech has come to be viewed as a form of speech that, for the most part, deserves constitutional protection as long as it does not present imminent and verifiable harm. Nonetheless, many technology, policy, and law experts have emphasized the need to evaluate deepfake technology in a new light due to its ability to diffuse rapidly through social media platforms (Citron and Chesney, 2018). In particular, scholars have pointed out the outdatedness of placing deepfakes in the same category of "false statements" described in previous court cases (Citron and Chesney, 2018; Chemerinsky, 2018; Manzi, 2019). Citron and Chesney (2018) illustrate how deepfakes' capacity to introduce unprecedented forms of exploitation, intimidation, and personal sabotage, as well as their ability to distort democratic discourse and manipulate public opinion, provide convincing reasons to consider whether the benefits of enabling deepfakes to circulate unchecked online outweigh the broad range of harm they pose to individuals and to society. Similar arguments have been made by researchers and academics who are increasingly wary of deepfakes' ability to produce multimedia content that is deliberately deceptive yet hardly distinguishable from its authentic counterpart. A Brookings Institution report suggests that, because deepfakes are so realistic, they can exert a considerable amount of influence over our "understanding of truth" (Villasenor, 2019). The report explains that by exploiting our inclination to trust the credibility of information we see with our own eyes, deepfakes can transform complete fiction into apparent reality or vice versa, resulting in a world where "truth itself becomes elusive, because we can no longer be sure of what is real and what is not" (Villasenor, 2019).

Referring to *NYT v. Sullivan*, Sunstein posits that the idea of democracy is a "double-edged sword" (Sunstein, 2019). The hypothetical scenario he provides involves a speaker intentionally lying about a politician and destroying her reputation in the process. In Sunstein's view, allowing this kind of harmful speech is not consistent with the idea that speakers need "breathing space" to preserve an "uninhibited, robust, and wide-open" system of free expression in which speakers and writers are not deterred by the prospect of lawsuits (Sunstein, 2019). Sunstein argues that unintentional mistakes, which may occur when one engages in open democratic debates, are fundamentally different from purposeful attempts to distort known facts. This further illuminates how deepfakes, which constitute a form of deliberate deception, would not be protected under *NYT vs. Sullivan* and that simply encouraging "more speech" as suggested by the marketplace of ideas is not the be-all and end-all when it comes to preserving a healthy democratic space for individuals to freely share their views with others (Waldman, 2019).

Chemerinsky (2018) also suggests that when the speech is false, the assumption that more speech is inherently better is less convincing; his reasoning can be summarized as follows: (1) Speech is protected partly because of the belief that the marketplace of ideas is the best way for truth to emerge. (2) The harmful effects of false speech *infect* the marketplace, and there is no reason to believe people will be able to discern facts from falsities. In accordance with his views, the advent of the internet and social media, which has enabled relatively unrestrained, low-cost capacity for communication of all kinds, has lent further credence to the apparent limitations of the assumption that "more speech" is inherently better.

The documented harms of deepfakes and their unique capacity to disseminate deliberately altered content to a sizable audience shed light on how deepfakes differ in their purpose and magnitude from the kinds of false speech the Court has come to protect through cases such as *NYT v. Sullivan* and *U.S. v. Alvarez*. At the most basic level, deepfakes are the embodiment of highly advanced forms of deception made possible by artificial intelligence technology in the digital age. Just by this fact alone, actual malice would not be difficult to prove, as long as the disseminators of the deepfakes in question are either the creators themselves or are fully aware that the material they are distributing are deepfakes. The component of demonstrating "legally cognizable harm" and showing that a causal link exists between the regulation of deepfakes and the prevention of injury , however, would be much more challenging, considering that the scope of harm deepfakes pose is not limited to individuals but covers a wide variety of audiences. In fact, the greater harm lies in deepfakes' ability to threaten democratic discourse altogether. Because the rise of social media, combined with the effects of confirmation bias, has rendered democratic processes especially vulnerable to misinformation and disinformation, it is important to recognize that the conventional rationale for protecting false speech may – rather than strengthening the marketplace of ideas – ironically undermine the kinds of uninhibited, robust, and wide-open interactions the marketplace of ideas serves to protect.

PART III: MALICIOUS DEEPFAKES VS. SATIRE AND PARODY – WHAT'S THE DIFFERENCE?

Although the harms of deepfakes are widely recognized and the notion that deepfakes should be protected as a form of false speech exposes serious legal limitations in addressing these harms, one needs to bear in mind that prohibiting the use of deepfakes altogether may risk censoring lawful and socially valuable speech, such as parodies and satires. Consider the following hypothetical scenario, which helps to shed light on the rationale behind the need to distinguish malicious deepfakes from deepfakes that may be used legally as a tool to satirize or parody certain aspects of society or the government (HG.org).

Let us assume that an avid supporter of Joe Biden, the leading Democratic candidate for the 2020 U.S. presidential election, decides to create a deepfake video of President Trump admitting to accusations against him as a way to *satirize* the fact that the President, despite much evidence in support of his involvement in numerous corruption scandals, has consistently denied his role. If this video were to circulate around social media under the guise of an authentic recording, that would amount to a legitimate case of libel, which is punishable by law. However, if the creator of this deepfake decides to clearly label the video as a deepfake in order to prevent its viewers from confusing the fake content as the truth, his or her malicious intent is substantially nullified. At that point, the question of liability for the deepfake and its potential negative consequences becomes less obvious.

Moreover, if we assume that the creator of the Trump deepfake leveraged the unique capacities of deepfake technology, not to deliberately deceive its viewers, but to simply increase the efficacy and reachability of his satirical message, limiting this kind of use becomes even more concerning. Although malicious deepfakes are fundamentally different from the kinds of false speech protected under *NYT v. Sullivan* and *U.S. v. Alvarez*, if deepfake technology were to be incorporated as a tool to satirize or parody the government, and the creator labels their work as a deepfake, the resulting deepfake would reasonably be protected as a socially valuable speech in the context of First Amendment law. The discrepancies that exist even within the category of deepfakes further illustrates the highly contextual nature of determining the limits of free speech protections, as suggested by Chereminsky (2018).

PART IV: POLICY IMPLICATIONS

As illustrated in Part III, indiscriminately prohibiting all forms of deepfakes may compromise socially valuable speech, such as parodies and satires. However, in response to growing concerns over the spread of misinformation ahead of the 2020 presidential election, many social media companies are implementing new policies to detect and remove content that has been deliberately altered and is likely to cause serious damage.

Notably, Twitter recently began applying a label to tweets containing synthetic or deliberately altered forms of media (Paul, 2020). The company also said it would actively remove any intentionally misleading manipulated content that is likely to cause harm (e.g. content that could cause violence, voter suppression, or privacy violations) (Paul, 2020). Moreover, Alphabet Inc's YouTube said it would remove any content that has been technically manipulated or doctored and may pose a "serious risk of egregious harm" (Paul, 2020). Similarly, Facebook announced its plan to ban certain manipulated photos and videos from its platform. It is important to note, however, that Facebook explicitly exempted content that is parody or satire from its new policy to combat the spread of deepfakes, further illustrating the need to clearly distinguish the use of deepfakes for the purpose of satire or parody from other malicious uses. For the most part, the actions that many social media companies are taking to combat the growing problem with deepfakes on the internet is promising as it brings awareness and, therefore, prudence to those who continue to seek robust ways to manage malicious deepfakes online.

However, social media companies' preventative policies are insufficient to safeguard democratic institutions from the spread of misinformation and disinformation in the digital age. In addition to the protections conferred by the First Amendment, the "fair use"[3] doctrine in copyright law, and section 230 of the Communications Decency Act (CDA)[4] provide ample room for malicious actors to continue infecting online communication networks with harmful deepfakes.

Some legal scholars have suggested that amending section 230 of CDA, which shields social media companies from liability for unlawful user-generated content as long as they take reasonable steps to prevent or address unlawful content posted by their users, would be a prudent way to comprehensively address the ways in which online content is published and distributed. Although amending section 230 of CDA so that social media companies are no longer immune from liability could certainly incentivize social media companies to discourage users from spreading misinformation and disinformation, Manzi (2018) suggests that this move alone would not eliminate the issue of fake news and deepfakes online. Manzi (2018) points to the actual malice standard as proof, in that while false speech that harms *specific* individuals would subject re-publishers to liability, false speech that causes *general* harm is currently unactionable.

---

[3] "Fair use" doctrine in copyright law is a legal doctrine that promotes freedom of expression by permitting the unlicensed use of copyright-protected works in certain circumstances. Section 107 of the Copyright Act provides the statutory framework for determining whether something is a fair use and identifies certain types of uses – such as criticism, comment, news reporting, teaching, scholarship, and research – as examples of activities that may qualify as fair use (U.S. Copyright Office, 2020).

[4] Section 230 of the Communications Decency Act, which was passed in 1996, says an "interactive computer service" can't be treated as the publisher or speaker of third-party content. This protects websites from lawsuits If a user posts something illegal, although there are exceptions for copyright violations, sex work-related material, and violations of federal criminal law (Newton, 2020).

Above all, the biggest obstacle to combating the spread of deepfakes seems to lie in the strong constitutional framework around false speech, especially false political speech, whose constitutional value within the marketplace of ideas framework is explicated in *NYT v. Sullivan*. Considering deepfakes' deliberately deceptive nature and the unique harms they pose to democratic processes, it is reasonable to believe that deepfakes should not be assessed using the same kind of rationale employed in *NYT v. Sullivan* and *U.S. v. Alvarez*. Nonetheless, until there is a serious reconsideration or perhaps a reversal of the Supreme Court decisions on the constitutionality of false speech, statutes criminalizing malicious and false speech would not be able to withstand the constitutional challenge rooted in the actual malice standard, as well as the *Alavarez* requirements that the speech causes "legally cognizable harm" and a direct causal link exists between the restriction imposed and the injury to be prevented" (NYT v. Sullivan; U.S. v. Alvarez).

CONCLUSION

Discussing the value of free speech in conjunction with the use of artificial intelligence technology such as deepfake technology is a relatively new phenomenon. As illustrated in the paper, there are numerous complexities and limitations involved in striking a balance between protecting the valuable uses of deepfake technology and regulating those that cause considerable harm not only to individuals but to society as a whole. In particular, the outdatedness of using the rationales articulated in *NYT v. Sullivan* and *U.S. v. Alvarez* regarding the value of false speech presents constitutional challenges against efforts to curb the threat of malicious deepfakes.

As difficult as this task may seem, the stakes of constructing a more comprehensive and robust legal framework are growing exponentially as technology continues to progress and reshape our societies. That said, a reconsideration of *NYT v. Sullivan* and *U.S. v. Alvarez* is necessary in order to distinguish deepfakes and other types of digitally manipulated content from the kind of false speech referred to in these cases. One way this can be done is by delineating the specific contexts in which deepfakes can be effectively regulated. In particular, the paper recommends that future research focus on how malicious deepfakes fundamentally differ from deepfakes that are used for socially valuable purposes and possibly come up with ways to further differentiate the myriad uses of deepfakes beyond the one presented in this research.

REFERENCES

Brownlee, J. (2019, July 19). A Gentle Introduction to Generative Adversarial Networks (GANs). Retrieved from https://machinelearningmastery.com/what-are-generative-adversarial-networks-gans/

Casad, B. J. (2019, October 9). Confirmation bias. Retrieved from https://www.britannica.com/science/confirmation-bias

Chesney, R., & Citron, D. K. (July 14, 2018). Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. *SSRN Electronic Journal*. doi: 10.2139/ssrn.3213954

Clement, J. (2020, April 1). Number of social media users worldwide 2010-2021. Retrieved from https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/

Fischer, W. (2019, October 10). California's governor signed new deepfake laws for politics and porn, but experts say they threaten free speech. Retrieved from https://www.businessinsider.com/california-deepfake-laws-politics-porn-free-speech-privacy-experts-2019-10

Fournier, A. (2020, April 24). The danger of disinformation in a time of crisis. Retrieved from https://thehill.com/opinion/cybersecurity/494455-the-danger-of-disinformation-in-a-time-of-crisis

Franklin, S. B. (2019, August 13). This Bill Hader Deepfake Video Is Amazing. It's Also Terrifying for Our Future. Retrieved from https://www.newamerica.org/oti/in-the-news/bill-hader-deepfake-video-amazing-its-also-terrifying-our-future/

Franks, M. A., & Waldman, A. E. (2019). Sex, Lies, and Videotape: Deep Fakes and Free Speech Delusions. *Maryland Law Review*, *78*(4), 892–898.

Gertz v. Robert Welch, Inc., 418 U.S. 323 (1974)

Ghaffary, S. (2020, February 4). Twitter is finally fighting back against deepfakes and other deceptive media. Retrieved from https://www.vox.com/recode/2020/2/4/21122653/twitter-policy-deepfakes-nancy-pelosi-biden-trump

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A.C., & Bengio, Y. (2014). Generative Adversarial Networks. *ArXiv, abs/1406.2661*.

Hall, H. K. (2018). Deepfake Videos: When Seeing Isn't Believing . *Catholic University Journal of Law and Technology*, *27*(1). Retrieved from https://scholarship.law.edu/jlt/vol27/iss1/4

HG.org. (n.d.). Retrieved from https://www.hg.org/legal-articles/freedom-of-speech-why-satire-is-protected-34438

Hustler Magazine, Inc. v. Falwell, 485 U.S. 46 (1988)

Franklin, S. B. (2019, August 13). This Bill Hader Deepfake Video Is Amazing. It's Also Terrifying for Our Future. Retrieved from https://www.newamerica.org/oti/in-the-news/bill-hader-deepfake-video-amazing-its-also-terrifying-our-future/

Libby, K. (2019, August 13). This Bill Hader Deepfake Video Is Amazing. It's Also Terrifying for Our Future. Retrieved from https://www.popularmechanics.com/technology/security/a28691128/deepfake-technology/

Mahadevan, A. (2020, April 27). Why Trump's comments on using disinfectants, sunlight to treat COVID-19 are wrong. Retrieved from https://www.poynter.org/fact-checking/2020/why-trumps-comments-on-using-disinfectants-sunlight-to-treat-covid-19-are-wrong/

Manzi, D. C. (2019). Managing the Misinformation Marketplace: The First Amendment and the Fight Against Fake News. *Fordham Law Review*, *87*(6). Retrieved from https://ir.lawnet.fordham.edu/flr/vol87/iss6/12

New York Times Co. v. Sullivan, 376 U.S. 254 (1964)

Panetta, G. (2020, April 24). Maryland said it's received 100 calls to its coronavirus hotline inquiring about Trump's suggestion that disinfectants might be able to treat COVID-19. Retrieved from https://www.businessinsider.com/maryland-hotline-got-100-calls-about-disinfectant-as-coronavirus-cure-2020-4

Paul, K. (2020, February 4). Twitter to label deepfakes and other deceptive media. Retrieved from https://www.reuters.com/article/us-twitter-security/twitter-to-label-deepfakes-and-other-deceptive-media-idUSKBN1ZY2OV

Petryna, A. (2013). *Life exposed: biological citizens after Chernobyl*. Princeton, NJ: Princeton University Press.

Priebe, M. (2020, April 23). The COVID-19 Infodemic: Combatting 'Dangerous' Misinformation on Social Media. Retrieved from

https://comartsci.msu.edu/about/newsroom/news/covid-19-infodemic-combatting-dangerous-misinformation-social-media

Romm T. (2020, January 7). Facebook bans deepfakes, but new policy may not cover controversial Pelosi video. Retrieved from https://www.washingtonpost.com/technology/2020/01/06/facebook-ban-deepfakes-sources-say-new-policy-may-not-cover-controversial-pelosi-video/

Rouse, M. (2018, June 25). What is deepfake (deep fake AI)? Retrieved from https://whatis.techtarget.com/definition/deepfake

Sample, I. (2020, January 13). What are deepfakes – and how can you spot them? Retrieved from https://www.theguardian.com/technology/2020/jan/13/what-are-deepfakes-and-how-can-you-spot-them

Sanders, A., & Sommerfeldt, C. (2020, April 26). A spike in New Yorkers ingesting household cleaners following Trump's controversial coronavirus comments. Retrieved from https://www.nydailynews.com/coronavirus/ny-coronavirus-new-yorkers-household-cleaners-trump-20200425-rnaqio5dyfeaxmthxx2vktqa5m-story.html

Schultz, D., & Hudson, D. L. (2017, June). Marketplace of Ideas. Retrieved from https://www.mtsu.edu/first-amendment/article/999/marketplace-of-ideas

Schwartz, O. (2018, November 12). You thought fake news was bad? Deep fakes are where truth goes to die. Retrieved from https://www.theguardian.com/technology/2018/nov/12/deep-fakes-fake-news-truth

Suciu, P. (2019, October 11). More Americans Are Getting Their News From Social Media. Retrieved from https://www.forbes.com/sites/petersuciu/2019/10/11/more-americans-are-getting-their-news-from-social-media/#7343dc4f3e17

Sunstein, C. (2019, February 22). Cass Sunstein: Clarence Thomas has a point about free-speech law. Retrieved from https://www.twincities.com/2019/02/25/cass-sunstein-clarence-thomas-has-a-point-about-free-speech-law/

Susan B. Anthony List v. Driehaus, 573 U.S. 149 (2014)

United States v. Alvarez, 567 U.S. 709 (2012)

Villasenor, J. (2019, February 14). Artificial intelligence, deepfakes, and the uncertain future of truth. Retrieved from

https://www.brookings.edu/blog/techtank/2019/02/14/artificial-intelligence-deepfakes-and-the-uncertain-future-of-truth/

Waldman, A. E. (2018). The Marketplace of Fake News. *Penn Law Journals*, *20*(4). Retrieved from https://scholarship.law.upenn.edu/jcl/vol20/iss4/3